High Capacity Relay Protocols for Wireless Networks

Yijia Fan, Ioannis Krikidis, Chao Wang, John S. Thompson, and H. Vincent Poor

(Invited Paper)

Abstract: **Over the last five years, relaying or multihop techniques have been intensively researched as means for potentially improving link performance of wireless networks. However, the data rates of relays are often limited because they cannot transmit and receive on the same frequency simultaneously. This limitation has come to the attention of researchers, and recently a number of relay techniques have been proposed specifically to improve the data efficiency of relaying protocols. This paper surveys transmission protocols that employ first single relays, then multiple relays and finally multiple antenna relays. A common feature of these techniques is that novel signal processing techniques are required in the relay network to support increased data rates. This paper presents results and discussion that highlight the advantages of these approaches.**

Index Terms: **Cooperative communications, diversity-multiplexing tradeoff, multiple input multiple output (MIMO), multiplexing gain, relaying, signal processing techniques, transmission protocols.**

I. INTRODUCTION

In recent years, cooperative diversity protocols (e.g., [1]–[5]) have been developed to improve link performance in wireless networks, where terminals assist each other by relaying transmissions. A simple cooperative network typically involves one source, one destination and one or more relay terminals which aid the source in communicating to the destination. However, one key limitation is the half-duplex constraint which arises because wireless relay terminals are not able to receive and transmit simultaneously at the same radio frequency. Thus, many protocols use two time (or frequency) slots to allow the source and then the relay to transmit their signals. In Rayleigh fading environments, these protocols can provide diversity gain to reduce the probability of deep fades and hence to improve link reliability [2], [4], [5]. However, when compared with direct transmission, relaying will become less spectrally efficient for high signal to noise ratio (SNR) values due to its inefficient use of two time slots to transmit one packet.

In this paper, we argue that this bandwidth inefficiency needs to be overcome in order to allow relaying techniques to be usefully employed over a wider range of operating SNRs. We begin in Section II by reviewing the conventional relaying protocol described above and highlight its limitations at high SNRs. A simple approach to try to minimize the half-duplex impairment is through the use of feedback from the destination. If the destination decodes the source transmission in the first time slot, it can send an acknowledgement (ACK) packet to the relay and source terminals. This knowledge can be used by the relay to avoid retransmitting the received data in the second time slot. Instead, the source can continue to transmit new information in that time slot, improving the communication data rate. One potential drawback of this scheme is that the throughput improvement of this scheme depends critically on the relative channel quality of the respective source-destination and source-relay(s)/relay(s)-destination links. If the source-destination SNR is much poorer than for the relay links, then the relay transmission in time slot two will be required most of the time.

This paper will focus on how signal processing techniques can be used to improve the data transport efficiency of conventional relaying protocols. A key idea is to make the throughput gains as robust as possible to variations in channel conditions on the links between source, relay and destination. The first approach to this problem is described in Section III of the paper and involves the use of non-orthogonal relaying protocols proposed in [3], which permit the source to transmit in the same time slots as the relay. This creates interference between the transmitters, but with the application of signal processing techniques to tackle the interference at the receiver, the overall efficiency of the source-destination link can be increased. If the relay decodes the source signal before transmitting, it can use a dynamic form of this protocol [6] where it is permitted to start retransmitting without necessarily needing to wait for the second time slot to begin.

An alternative way to improve link throughput is through the use of multiple antennas, either located at one terminal, or distributed across multiple nodes as described in Section IV. The concurrent relaying approach [7], [8] uses two relays to overcome the half-duplex limitations in relaying. The idea is that at any time one relay receives information from the source while the other relay is transmitting to the destination. Extra signal processing is needed at both the relay and the destination to tackle interference. However, this approach provides the diversity gain benefits of the conventional protocols, while avoiding the associated rate loss at high SNRs. Some authors have also considered how relaying techniques can be used with terminals using multiple antennas, as discussed in Section V. In this case, multiple input multiple output (MIMO) links are formed between each pair of terminals. An interesting new relaying ap-

Manuscript received March 26, 2009.

Y. Fan's and H. V. Poor's work was supported in part by the U.S. National Science Foundation under Grants ANI-03-38807 and CNS-06-25637. I. Krikidis', C. Wang's and J. S. Thompson's work has been funded by EPSRC and by the Industrial Companies who are Members of Mobile VCE and Scottish Funding Council for the Joint Research Institute with the Heriot-Watt University which is a part of the Edinburgh Research Partnership.

Y. Fan and H. V. Poor are with the Department of Electrical Engineering, Princeton University, Princeton, NJ, 08544, USA, email: {yijiafan, poor}@princeton.edu.

I. Krikidis, C. Wang, and J. S. Thompson are with the Institute for Digital Communications, University of Edinburgh, Edinburgh, EH9 3JL, UK, email: {i.krikidis, chao.wang, john.thompson}@ed.ak.uk.

proach known as filter-and-forward has been proposed to minimize the relay complexity in this scenario. This paper will discuss the major protocols that have been described in Section VI and finally present conclusions in Section VII.

II. RELAY SYSTEM MODEL

In this section we will define the basic relaying setup that will be discussed throughout the paper. The conventional relay system model assumes the presence of one source (S), one relay (R), and one destination (D). When the source transmits, the network can exploit the fact that the transmission is overheard both by the relay and the destination. The most widely studied transmission protocol [1], [2] can be described as follows: for the first time slot, the source broadcasts the message to both the relay and the destination as shown in Fig. 1(a); the relay then forwards its received signals to the destination in the second time slot to assist the destination to decode, as shown in Fig. 1(b). We term these protocols as "conventional protocol" and they may be used in ad hoc networks or as part of a cellular network using either fixed relays or mobile terminals.

A. Amplify or Decode at the Relay

Most work in this area considers one of two main ways for the relay to process its signal. In *amplify-and-forward* (AF) relaying, the relay simply demodulates its received signal to baseband, records and then amplifies and retransmits this signal at a later time, as shown in Fig. 1(c). The advantage of this approach is that minimal signal processing is required at the relay for the message to be conveyed to the destination. However, this approach will not remove the effects of receiver noise and interference from the retransmitted signal. In *decode-and-forward* (DF) relaying, the relay completely decodes the data from the source to recover the original source message. This data can then be reencoded and sent to the destination as shown in Fig. 1(d). This arrangement has the advantage that receiver noise can be removed from the signal if it is decoded correctly. However, the clear disadvantage of DF relaying is the higher processing complexity that is required of the relay terminal. Some hybrid schemes have been proposed including *decode-amplify-forward* (DAF) [9]. This scheme uses DF when the source-relay link SNR is sufficiently good for the relay to decode the source message without errors. Otherwise AF is used to forward information to the destination. Such hybrid schemes can improve performance over using only DF or AF in isolation, but increase the complexity of relay processing.

B. Diversity Performance of Relays

Using either AF or DF relaying, the relay can convey its received signal to the destination. The destination can then combine both the signal it received from the source and that received from the relay to estimate the transmitted data. As these signals are received over two independent channels, the destination can exploit *diversity* to improve performance in the presence of multipath propagation from transmitter to receiver. With multipath, there is a significant possibility of destructive addition of signals propagating along these paths, leading to very low levels of received power at the receiver. This effect is called fading and can

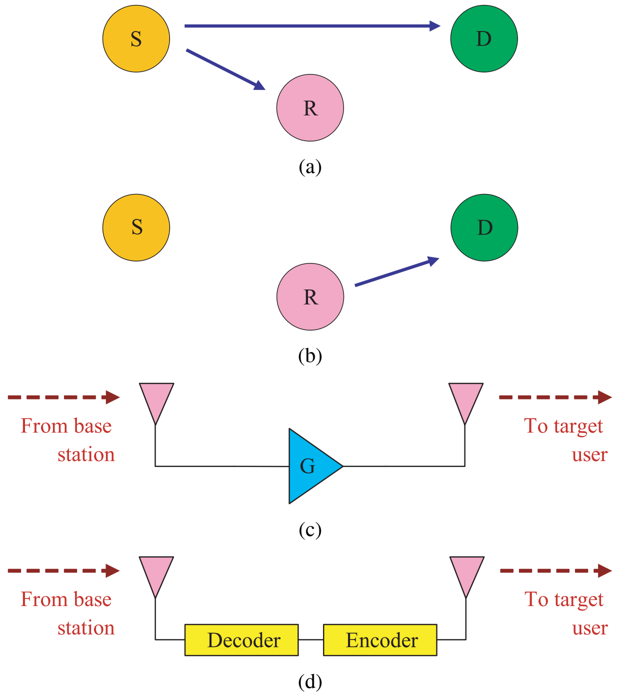


Fig. 1. (a) Source S transmits to the relay R and destination D in time slot 1 of the conventional relay protocol, (b) relay R transmits to the destination D in time slot 2, (c) block diagram of an AF relay with amplifier gain G, and (d) block diagram of a DF relay.

lead to bursts of packet errors when it occurs. In the presence of two independent channels, i.e., S-D and S-R-D, the likelihood of both being simultaneously in a deep fade is much less than for the S-D path alone.

In addition to diversity gain, relay protocols can also provide significant link gain when the relay is located between the source and destination. The S-R and R-D links experience much lower attenuation than the S-D link so that relaying can also be used to extend the coverage of wireless networks. However, relays cannot be usefully exploited in all SNR conditions because of the half-duplex nature of relays. As the SNR increases and fading conditions become less severe, relay links become limited by the requirement for two time slots to forward information. Ultimately, the SNR of the direct S-D link can become so high that there is no advantage in using relaying. Instead the two relaying time slots are better used to to send two data packets on the direct link without the assistance of the relay. We will refer to this limitation as the *multiplexing loss* of conventional relay protocols at high SNR values.

As shown in [10] there is a trade-off between diversity and multiplexing in relay networks. That paper considers *full-duplex* relay systems where relays can receive and transmit simultaneously. Even in this case, however, at infinite SNRs full-duplex relay networks cannot outperform the direct link in terms of multiplexing efficiency. This result provides some idea of the potential gains that we can seek to obtain through the protocols described here. Before moving on to investigate techniques that exploit signal processing to improve performance, we start by discussing simple techniques that use ACK-based feedback signals from the destination.

C. Incremental Relaying and ARQ-Protocols

Perhaps the simplest modification to the conventional protocol is to exploit ACK packets that are typically sent from the destination to the transmitter to indicate the success or failure of packet decoding. In the conventional protocol, if the destination is able to decode its received packet after the first of the two time slots, Fig. 1(a), it can transmit an ACK packet on the backwards channel to the relay and the source. This allows the relay to avoid transmitting data in the second slot in favour of the source transmitting an additional data packet, as shown in Fig. 1(a). If the source does not decode the signal, a negative ACK (NACK) is transmitted by the destination, so that the relay will forward its signal to the destination in the second time slot, Fig. 1(b).

The first protocol to exploit such feedback was presented in [2] and was called incremental relaying. It was combined with AF relaying techniques to ensure that the relay can always transmit its signal in the event that the destination transmits a NACK packet. The probability with which the destination can decode the signal will dictate the achievable throughput. The higher the probability, the less frequently relay operation is required and the higher the overall data throughput from source to destination. Thus throughput depends critically on the quality of the source-destination link and on the ability of the relay to aid the destination in decoding its signal. More recent work has proposed the use of hybrid automatic repeat request (ARQ) transmissions from both the source and relay terminals to aid the destination in decoding the source signal [11]. Hybrid ARQ involves the use of different data encoders to those used by the source to transmit additional information to help the destination decode the source information.

As noted in the introduction, the gain in throughput by using ACK/NACK feedback depends strongly on the relative SNRs of the source-destination and the relay links. If the relay provides a much more reliable link than the direct source-destination link for the required data rate, then feedback will not provide much improvement in throughput. Any gain in throughput can be exploited in ad hoc peer-to-peer networks, in which the spectrum resources are managed in a decentralized fashion. If the transmission of the ACK packet avoids the relay retransmitting the source data in the second time slot, then that time slot can be used directly by the source to transmit new data, increasing the spectral efficiency of that link. However, it may be more difficult to use this approach in cellular wireless networks, where the network has to allocate frequency resources to source and relay terminals. In this case, the network may allocate some of its bandwidth resource to source transmissions and the remainder to relay transmissions. In particular, for the downlink (base-to-terminal link), it is important to avoid interference from high power base station transmissions to lower power relay transmissions [12]. In this scenario, if the relay does not transmit, the base station cannot reuse the relay's time slot to transmit new data. However, the fact that the relay does not transmit will reduce interference to other relay links operating nearby.

We will now move on to discuss alternative approaches to ACK/NACK feedback, which rely instead on signal processing techniques at relays or destinations to mitigate interference. This approaches have already been described briefly in the Introduction, but will now be discussed in more detail. We begin our discussion of efficient relaying protocols by studying non-orthogonal relay protocols and in particular the non-orthogonal amplify-and-forward (NAF) method.

III. IMPROVING MULTIPLEXING GAIN IN SINGLE RELAY NETWORKS

In this section, we present some cooperative techniques that overcome the multiplexing loss of the half-duplex constraint and improve system performance for the case of one relay. The system model considered here follows that of Section II. In contrast with previous work, e.g., [1]–[3], which assumed that the channel gains do not change during the transmission of a cooperation frame (i.e., two data slots) here we suppose a classical quasi-static channel which remains constant during the transmission of one time slot but changes independently from one time slot to another. This system assumption corresponds to relative motion between the source, relay and destination over the communication interval. Furthermore, although the presented approaches are based here on the AF cooperative policy in order to reduce relay complexity, they can also be applied with some modification to DF schemes. The dynamic DF (DDF) protocol in [6] operates such that the relay is allowed to start transmitting to the destination as soon as it has decoded the source signal. However, the relay needs to encode its waveform in a different way from the source in to allow the destination to decode both S and R signals efficiently.

A. Non-Orthogonal AF (NAF) Protocol

The NAF protocol was proposed by Nabar *et. al.* [3] and has been proved to be the optimal AF scheme for a half-duplex single-relay channel by Azarian *et. al.* [6] in terms of providing the best tradeoff between diversity gain and multiplexing gain. It is an optimization of the conventional AF protocol, termed orthogonal AF (OAF) in [2] in which two time slots separate the source and the relay transmissions. More specifically, in order to overcome the data rate loss arising from the inactivity of the source during the cooperative channel in OAF, the NAF protocol allows the source to be active during the relaying transmission. Its superiority comes from the fact that the source keeps transmitting new data ($x(t)$) while the relay forwards previously "overheard" data ($x(t-1)$) with $x(t) \neq x(t-1)$. The destination experiences inter-symbol interference due to the simultaneous transmission of both $x(t)$ and $x(t-1)$.

The optimal method to decode this interference signal is *maximum likelihood (ML)* decoding [3], [8]. It compares all possible noise free received signals with the actual noisy one to find the most likely transmit signals. Unfortunately, the complexity of ML increases rapidly with the number of transmitters and the digital constellation size. A simpler suboptimal alternative is *successive interference cancellation* (SIC), which decodes the interfering signals in turn [8]. Once a particular signal has been decoded, it may be cancelled or subtracted out from the received signal to remove its effect on the receiver. The detection order is often chosen to detect the signal with maximum SNR at each stage, in order to avoid decision errors that will cause increased levels of interference. Compared to OAF, the required signal processing at the destination has clearly increased. The NAF protocol improves multiplexing gain as the source continuously transmits new data to the destination. However, for a single-relay setting, it provides relaying for half of the time and therefore data are not always "protected" by diversity. The NAF protocol can also be generalized to the case when multiple relays are available, as explained in [6], [13].

B. Block-Fading Non-Orthogonal AF (BFNAF)

The protocol proposed here is similar to the classical NAF except that the source retransmits the same packet during the cooperative slot in [14]. The principal motivation behind this protocol is to add S and R transmissions coherently at the receiver in the second slot to improve received signal power. The physical link between the source and the destination changes between the two time slots of the protocol due to the slot-based block-fading nature of the channel. The source retransmission of the same data via another independent channel can also improve the system reliability compared to direct transmission. As we will show in the following discussion, this new NAF behavior is interesting for low data rates where the diversity gain is more important than the multiplexing-gain. Using the previous formulation of the conventional NAF scheme, the BFNAF protocol is characterized by the property $x(t) = x(t-1)$ for the tth cooperative frame.

This choice of waveform for $x(t)$ means that the BFNAF technique can yield low-complexity implementations. This is in contrast to other NAF-based protocols which require advanced signal processing techniques at both relay and destination [6], [13]. We assume that the required signal processing is limited to a simple cophasing operation of the simultaneous transmissions (S→D, R→D) to allow constructive addition of signal power at the receiver [11]. This means that the two time slot signals can be combined using maximal ratio combining followed by standard demodulation procedures at the receiver [14]. Fig. 2 schematically presents the transmission structure of the considered AF protocols.

B.1 Optimal Power Allocation for NAF Protocols

A basic problem introduced by the above NAF schemes is how to select transmit power levels in the network. If we assume that the total transmit power in the network is fixed, in order to avoid excessive interference to other nodes, one can determine how it should be distributed among the source and relay transmissions to optimize performance. In the studies described in [14]–[16] the optimization criterion is the *outage probability* which is the likelihood that the system cannot deliver data at a given rate without errors. In order to avoid excessive signalling of channel values between the nodes, the optimization depends only on the average channel statistics. Given the complexity of analyzing the NAF and BFNAF protocols, simplified bounds have been been used in [14] to determine the optimal power values. The proposed bounds suppose the two simultaneous transmissions of the second slot are orthogonal (i.e., different frequencies or spread spectrum codes). This assumption exploits all the diversity degrees of the channel model and efficiently approximates the optimal power allocation without complicated computations.

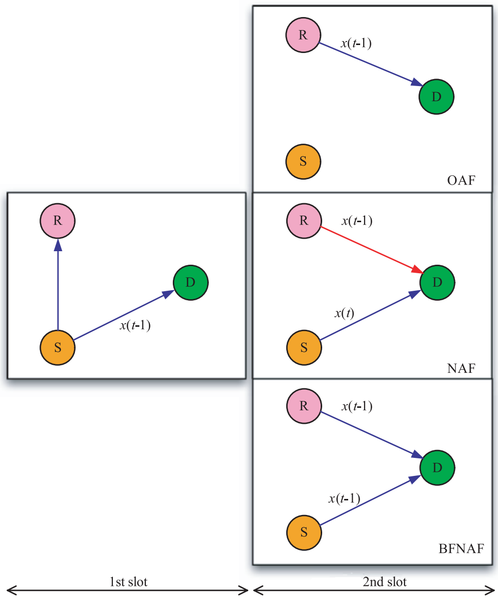


Fig. 2. The transmission structure of the considered AF protocols: OAF, NAF, and BFNAF.

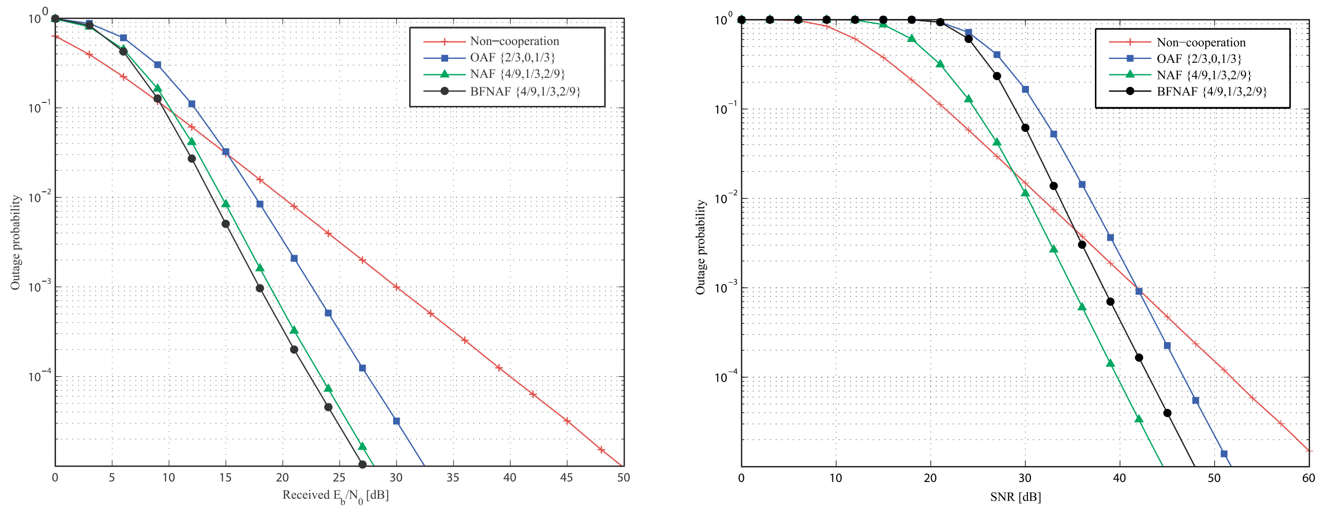
B.2 The ARQ Case

The above AF protocols straightforwardly can be extended for the case of an ARQ scheme as discussed above in Section II. In this case cooperation is used only when the direct link is not able to deliver correctly the data at the destination [17], [18]. The switching criterion between the two modes of the protocol (direct or cooperative) is the instantaneous quality of the direct link and more specifically its outage behavior. When the direct link is "good," which in terms of outage translates to an instantaneous capacity higher than the required spectral efficiency, the information can be transmitted directly to the destination and thus the selected mode is the direct non-cooperation transmission. On the other hand, in the case that the instantaneous direct link is characterized as "bad," which means a capacity lower than the required rate, the selected mode is cooperative (OAF, NAF, and BFNAF).

C. Numerical Results

In this section, we illustrate the performance of the above protocols by means of Monte-Carlo simulations. By default, we consider a three-node symmetric network, where all the channels coefficients are independent and identically-distributed (i.i.d.) Rayleigh random variables with unit variances. The selected performance metric is the outage probability for a given information rate R_0 measured in bits per channel use (BPCU).

In Fig. 3, we compare direct transmission, OAF, NAF, and BFNAF using the optimal power allocation policy described above for two data rate values. Fig. 3(a) shows results for a low desired data rate $R_0 = 1$ BPCU and Fig. 3(b) is for a higher rate $R_0 = 4$ BPCU. The numbers in brackets represent the normalized power allocation for source transmissions in the two time slots, then the relay transmission. As can be seen, the NAF protocols significantly outperform the conventional OAF scheme. The multiplexing loss of the OAF scheme results in poor performance for both cases. Finally, the most important observa-



(a) (b)

Fig. 3. Outage probabilities for conventional protocols; non-cooperative, OAF, NAF, and BFNAF with optimal power allocation. The considered information rates are (a) 1 and (b) 4 BPCU; the brackets include the power allocation for the links $\{S \rightarrow D$ (slot-1), $R \rightarrow D$, $S \rightarrow D$ (slot-2)$\}$, respectively (normalization to unity).

tion is the comparison between the two NAF-based schemes. As we can see, the BFNAF protocol outperforms NAF at low spectral efficiencies (1 BPCU) but is outperformed by NAF at higher spectral efficiencies (4 BPCU). For example at 10^{-3} outage probability, the gain of BFNAF is equal to 1 dB for $R_0 = 1$ BPCU, while for $R_0 = 4$ BPCU, the gain of NAF is 3 dB. This result shows that maximizing signal power via the proposed BFNAF scheme is useful for low data rates where diversity gain is more important than the multiplexing gain available from NAF. The importance of multiplexing gain in NAF is visible for higher data rates. The performance improvement for NAF over BFNAF will translate directly into higher link capacity results for the NAF protocol. This observation led to a hybrid protocol in [14] that switches between NAF and BFNAF depending on the data rate.

Fig. 4 compares the ARQ version of the considered NAF-based schemes for a spectral efficiency equals to 1 BPCU. The performance of the non-cooperative scheme and OAF are given as references curves. In order to make a fair comparison, we note that an optimal power allocation is considered for all the schemes. First of all, we see the hybrid protocols overcome the problem that conventional protocols have in low SNRs. Due to the selected activation of the cooperative mode, the cooperative schemes are never outperformed by the non-cooperative case. Moreover, this selected activation improves the performance in comparison with the non-adaptive schemes (Fig. 3(a)). The hybrid version of the cooperative protocols provides the expected gains at low SNRs but also improves the performance in the high SNR regime.

IV. MULTIPLE SOURCE MULTIPLE RELAY PROTOCOLS

In this section, we will discuss spectrally efficient concurrent relay protocols in which multiple terminals are assisted by multiple relays to communicate to the destination.

A. Concurrent DF Relaying

The NAF protocol cannot easily be extended to DF relays as the diversity gain of the network is not improved in this case [19]. Instead, one may consider how multiple relays could be exploited to improve performance. In some papers, multiple relays are used to improve only diversity gain over single-relay systems. For orthogonal DF schemes, a higher diversity gain is accompanied by a even lower multiplexing gain [1]. This problem can be partly solved by requiring all the relays to utilize distributed space-time codes to transmit simultaneously [1] or using only the best relay to transmit during the second time slot [20], [21]. However, the half-duplex constraint still limits performance in these schemes.

Instead, for multiple-relay scenarios, the concept of successive relaying (independently proposed by [7], [8], and [13] in different contexts) is an effective approach to improving multiplexing gain over that of conventional protocols. The basic idea behind successive relaying is that two (or more) relays take turns helping the source so that the degrees of the freedom of the channel are efficiently used. For DF relaying, the single-source network studied in [8] is further extended to a multiple-source network in [22] (where it is termed concurrent DF relaying).

For example, the concurrent DF relaying protocol considers a five-node network with two sources S_1 and S_2, two half-duplex DF relays R_1 and R_2, and one common destination D, as displayed in Fig. 5. The transmitted messages from each source are divided into different frames, each containing L independent codewords.

For conventional time division multiple access (TDMA) direct source-destination transmission, the L codewords from each of the two sources are transmitted to the destination using $2L$ time slots, while for the conventional protocol with one relay [2], the transmission process must use $4L$ time slots due to the half-duplex operation of the relay. For concurrent DF relaying, one source and one relay are both allowed to transmit simultaneously. In the first time slot, Fig. 5(a), only S_1 transmits to R_1. At

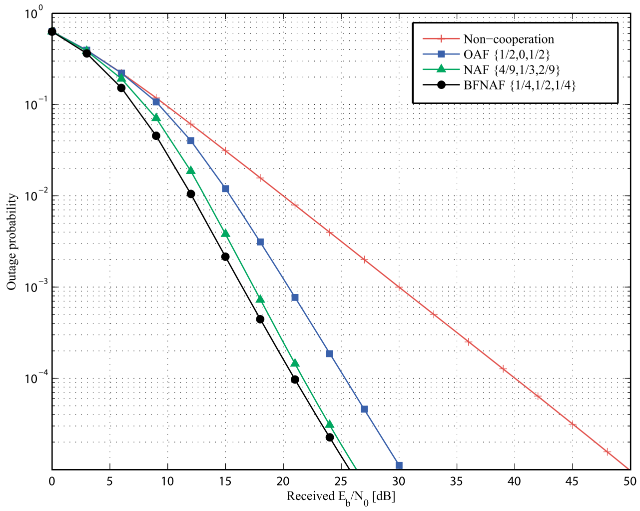


Fig. 4. Outage probabilities for ARQ-based protocols; non-cooperative, OAF, NAF, and BFNAF with optimal power allocation. The considered information rate is 1 BPCU; the brackets include the power allocation for the links $\{S \to D$ (slot-1), $R \to D$, $S \to D$ (slot-2)$\}$, respectively (normalization to unity).

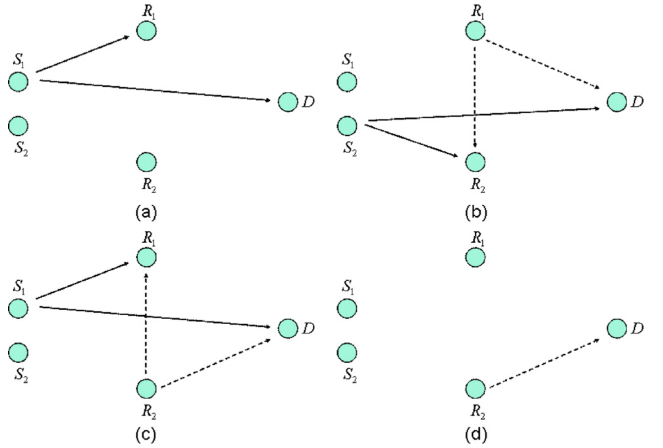


Fig. 5. Operation of the concurrent DF protoocl: (a) S_1 transmits in time slot 1, (b) S_2 and R_1 transmit in time slot i, (c) S_1 and R_2 transmit in time slot $(i+1)$, and (d) R_2 transmits in the final time slot (number $2L+1$).

any later time, when S_2 transmits, R_1 sends its received message from S_1, Fig. 5(b). This pattern alternates with S_1 transmitting simultaneously with R_2, Fig. 5(c). Finally, the transmission is completed by R_2 sending its final transmission, Fig. 5(d).

The major issue with concurrent DF relaying is the interference generated among relays when one relay is listening to its associated source, while the other relay is forwarding its source message to the destination. To suppress the interference, a simple decoding criterion is considered for each relay: If the interference between relays is stronger than the desired signal, the relay decodes the interference signal and subtracts it from the received signal before decoding the desired signal. Otherwise, the relay decodes the desired signal directly while treating the interference as additional noise.

Assuming that relays can always successfully decode their source messages, concurrent DF relaying mimics a $2L$ user multiple access MIMO channel except that the dimensions of signals are expanded in the time domain rather than space domain. Signal processing techniques for conventional point-to-point MIMO systems such as ML or SIC can thus be used directly. Since $2L$ codewords are transmitted to the destination using only $2L+1$ time slots, rather than $4L$ for the conventional protocol, the multiplexing gain performance of the system is significantly improved. Furthermore, when the frame length L is large, the multiplexing gain performance approaches that of the full-duplex relay bound of [10], which is the same as TDMA direct transmission. On the other hand, concurrent DF relaying also obtains diversity gain improvement over TDMA direct transmission because the destination receives duplicates of the source codewords from both direct and relay links. As in Section III and following the idea described in [23], this scheme can be simply combined with ACK/NACK feedback, by stopping the transmission protocol shown in Fig. 5 as soon as the destination decodes the transmitted signals and transmits an ACK packet. The sources are then free to start transmitting new data sequences. A similar approach has been analysed in [24], which shows that selecting relays to transmit/not transmit based on the channel conditions does not affect the overall diversity order of data detection at the destination. The use of different encoders at the relay compared to the source as in hybrid ARQ [19] may even permit relay networks to achieve similar capacity gains to MIMO systems, as described in [25], [32]. Therefore, compared with the conventional protocol, concurrent DF relaying makes relaying more beneficial.

Fig. 6 shows comparisons of direct transmission, the conventional DF protocol and the concurrent relaying protocol in terms of outage probability. Fig. 6(a) of the figure shows results for data rate $R_0 = 1$ BPCU, and it shows that at this data rate the relaying protocols significantly outperform direct transmission. The results also show the performance advantages of the concurrent DF relaying protocol which increase as L is changed from 1 to 2 time slots. Fig. 6(b) shows the performance comparison for $R_0 = 4$ BPCU which shows that the conventional DF protocol degrades significantly with respect to both concurrent DF relaying and direct transmission. This result highlights the fact that concurrent DF achieves a higher multiplexing efficiency [22] than the conventional DF protocol. The concurrent protocol also provides an improved diversity gain, visible from the steeper slope of the outage curve, when compared to direct transmission. For $L = 2$, the concurrent protocol offers the best performance at outage probability values below 0.1. Detailed analyses of the performance tradeoffs for successive and concurrent DF relaying can be found in [7], [8], [13], [22]. In the next section, we will discuss how relaying can be applied in configurations where each terminal has multiple antennas, rather than a single antenna.

V. APPLICATION OF RELAYING TO MIMO CHANNELS

For multiple-antenna or MIMO relay networks in which every terminal in the network is deployed with multiple antennas, studies have mainly concentrated on spatial multiplexing systems in terms of capacity or throughput analysis. Capacity

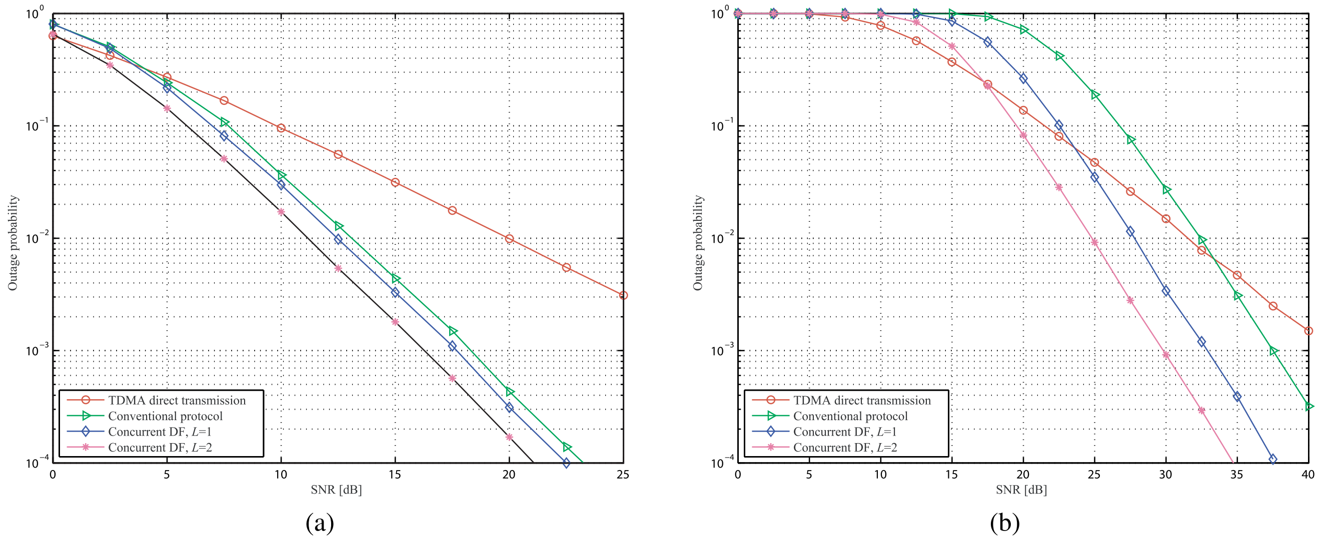


Fig. 6. Outage probabilities for direct transmission, the conventional DF protocol and the concurrent DF protocol using $L = 1$ and $L = 2$; part is (a) for rate $R_0 = 1$ BPCU and (b) for $R_0 = 4$ BPCU.

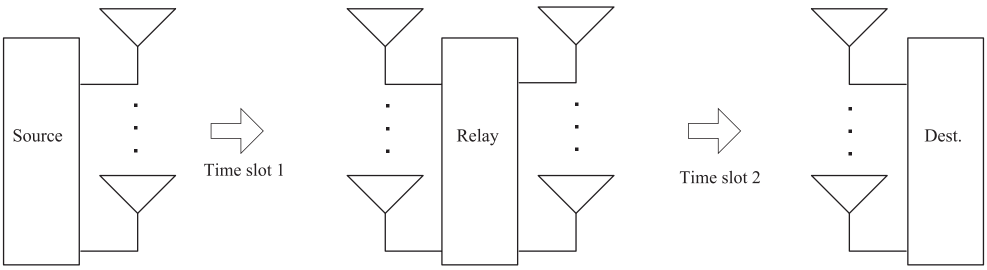


Fig. 7. Basic system model of a MIMO two hop relay network.

bounds for single relay MIMO channels are presented in [26]. While this work is primarily focused on the theoretical perspective, some more practical results considering specific signal processing algorithms and achievable rates have also been presented [27], [28].

A. System Model

A simple MIMO relay channel is shown in Fig. 7, in which the source is equipped with N_s antennas, the relay is equipped with N_r antennas and the destination is equipped with N_d antennas. Unlike Section II we assume here that the direct link is ignored due to shorter distance between the source and the relays, relative to the destination. This simplifies the receiver processing. However, note that when the direct link is strong, direct transmission is usually the preferred choice especially for a MIMO link. We shall now discuss MIMO relay configurations using both AF and DF relays.

B. Conventional Relaying Modes

For DF relaying, the relay first uses all of its N_r antennas to jointly decode the signals; it then demultiplexes the decoded message to form N_r data streams and uses all N_r antennas to re-encode and retransmit the data streams to the destination. The overall S-D capacity is typically constrained by the worse capacity of the S-R and R-D links. Since both of these links are likely to be strong compared with the direct link, the system capacity can be greatly improved. If the relay decodes the source transmission incorrectly, the relay obviously cannot send the correct information to the destination. However, if the transmission rate is below the end-to-end S-D capacity, the error rate can be made arbitrarily low by using powerful coding schemes such as low density parity check codes.

One practical disadvantage for DF relaying is its high complexity due to the multi-antenna transceiver structure, as ML decoding might be required at both the relay and the destination to achieve the available link capacity. Suboptimal schemes, such as SIC, can be used instead to reduce the receiver complexity. Even still, the processing complexity of SIC increases at least linearly with the number of transmit and receive antennas.

To overcome the relay processing complexity, AF relaying can be used instead. The process is much simpler as decoding or encoding is not required at the relay. However, one obvious defect for AF relaying is that the relay amplifies the receiver noise and thus causes a decrease in the SNR at the destination, which can be significant especially when the number of antennas at the relay is large (e.g., in the scenario where a fixed relay with large antenna array size is used).

C. Filter-and-Forward Relaying for MIMO Terminals

Unlike the situation for single antenna systems, in MIMO configurations the multiple propagation paths that exist between the transmit and receive antennas can be exploited to improve the relay's signal quality. We begin by discussing the AF relaying (AFR) protocol for MIMO systems. In order to improve the performance of AFR, the received signals at the relay antennas can be jointly filtered before AFR is used. Several recent papers have studied the optimal filter that should be used in the relay to maximize the end-to-end data capacity [27], [28], provided that R-D channel state information (CSI) is known to the relay. Overall, the strategy for selecting the optimal filter follows two steps:

(a) Decomposing the MIMO relay channel into several orthogonal, parallel spatial channels to avoid co-channel interference between different transmit antennas. This can be achieved by using the singular value decomposition to decompose the S-R and R-D channels into their constituent eigenvalues and eigenvectors.

(b) Allocating the transmit power optimally across the antennas at the relay (and the source, if the forward channel knowledge is available at the source).

While step (b) is usually complicated to perform with no apparent closed form solution, step (a) is simple to perform. It has been proved [27], [28], [31] that the optimal filter depends on the left singular vectors for the S-R channel and the right singular vectors for the R-D channel. These vectors are scaled by constants that determine the power allocation values across all the antennas at the relay. This design allows the end-to-end MIMO relay channel to be decomposed into several orthogonal, parallel relay channels. This further allows the signal to interference-plus-noise ratio (SINR) for each sub-channel to be optimized. Thus *filter-and-forward relaying* (FFR) offers a significant performance advantage compared with the AFR method in terms of capacity, particularly when the number of relay antennas N_r is higher than for the source N_s or destination N_d. The FFR scheme also simplifies decoding at the destination, since the orthogonal relay channels can easily be recovered by filtering. This avoids using high complexity ML or SIC decoding at both the relay and the destination.

FFR offers similar (sometimes even better [28]) performance to MIMO DF relaying (DFR). In DFR, as the relay decodes the data, it can demultiplex and remultiplex its received packets into a different set of transmit signals for the R-D link, in order to fully exploit the MIMO capacity available on that link.

Fig. 8 gives a simulation example taken from [27] in which the source and the destination are both equipped with 2 antennas and the relay is equipped with 8 antennas, on assuming the R-D CSI is available at the relay. Here we use average capacity (throughput), instead of outage probability, as a metric to measure the system performance. The average capacity is defined as the system capacity when the value of the SINRs at the receiver is available at the transmitter through certain feedback. Because the main purpose to use MIMO spatial multiplexing relay structure is purely to increase spectral efficiency, i.e., transmission rate. Therefore, we believe average throughput is a better metric than outage probability, which is primarily for outage (diversity) measure. In order to achieve a higher multiplexing rate, feedback (e.g., ARQ as one simplest example) is always needed in order for the system to adjust its rate. In this respect, the average capacity can be considered as the performance upper bound for all such kind of feedback schemes. Scheme AFR denotes AFR; FFR with equal power allocation denotes the FFR with the same power allocated to each parallel relay channel; Optimal FFR denotes the FFR with optimal power allocation at the relay; and scheme DFR denotes DFR. It can be seen that FFR schemes offer the similar performance to DFR, and significantly outperform AFR.

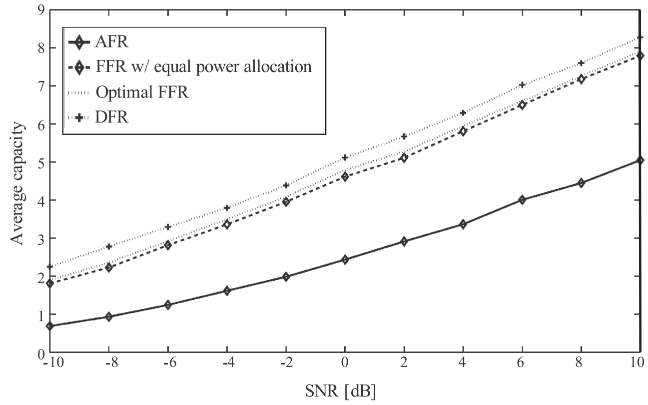


Fig. 8. Average capacity of single MIMO relay channels as a function of SNR when $N_s = N_d = 2$, and $N_r = 8$ (after [27]).

In terms of complexity, FFR is slightly more complicated than AFR, as it has to decode a training sequence (or feedback) to obtain the CSI at the relay, and needs a filter at the relay to refine the message. However, FFR is much simpler than DFR, as it avoids the need for a complicated ML/SIC decoding and encoding process at the relay. Both AFR and FFR instead require the use of ML/SIC decoding at the destination, a requirement which DFR can avoid provided the relay-destination CSI is available to the relay. The extension of the filter matrix design to consider both the direct and relay links can be found in [28]. Some discussion on the scenario where no forward CSI is available at the relay can be found in [27].

Hybrid ARQ schemes can be applied to the case of MIMO relay networks in a similar way to that described in Section II. In [27], the structure of the FFR algorithm lends itself to efficient ARQ schemes, since different parts of the data are transmitted on different orthogonal spatial channels. Only the data associated with those spatial channels where decoding was not possible need to be retransmitted [29], possibly using hybrid ARQ techniques. In Fig. 7, the direct source-destination link was neglected. However, [30] considers the case where data is transmitted on the direct source-destination link. A cooperative relay is only used to forward hybrid ARQ transmissions to the destination in case that decoding was not possible on the direct link.

VI. DISCUSSION

In order to put the different techniques into perspective and understand their requirements, Table 1 compares various protocols that we have discussed in this paper. We have considered whether relays require CSI for the R-D link; whether ML/SIC decoding is performed at the destination to mitigate interference; whether the relay decodes and re-encodes packets and finally whether the relay needs to perform demultiplexing/remultiplexing of data streams to optimize performance.

Table 1. Comparison of complexity and CSI requirements for different relaying methods. * denotes the fact that ML/SIC detection may need to be performed at the relay rather than the destination.

Algorithm	Forward CSI	ML or SIC decoding	Re-encoding	Demux	Section
OAF	No	No	No	No	III
NAF	No	Yes	No	No	III
BFNAF	Yes	No	No	No	III
con DF	No	Yes	Yes	No	IV
AFR	No	Yes	No	No	V
FFR	Yes	Yes	No	No	V
DFR	Yes	Yes*	Yes	Yes	V

Table 1 clearly shows that the NAF, concurrent DF, AFR, FFR, and DFR protocols require ML/SIC decoding at the destination to overcome interference generated by the protocol. The BFNAF and DFR methods exploit forward R-D CSI in order to optimize performance and avoid ML/SIC decoding at the receiver. The requirement for accurate CSI knowledge at the relay, combined with the requirement for extra signal processing to compute singular vectors in order to exploit this knowledge means that the processing burden is shifted from the destination to the relay. The concurrent DF and DFR methods are the only DF protocols considered in this paper. Finally, the DFR method is the only one that may require demultiplexing and remultiplexing of the received data at the relay, in order to best exploit the available data capacity on the R-D MIMO link.

In addition to the protocols discussed in this paper, there are a number of other interesting new research directions. Researchers are beginning to study *compress-and-forward* relaying [10], which can be viewed as a generalization of amplify-and-forward relaying. In this approach, source coding techniques are used in the relay to reduce redundancy in the relay's retransmitted signal. Recent research has also reconsidered the full-duplex multiplexing bound discussed in [10]. It turns out that it is sometimes possible to form a MIMO system using only a single antenna source but with multiple antennas available at the relay and the destination [25], [32]. This procedure works only at finite SNR values and depends on the S-R distance being much less than the S-D distance. It also uses a different encoder in the relay than at the source, similarly to the DDF protocol mentioned in Section III. There is also interest in how a single relay can be shared among multiple sources communicating to one destination (the multiple access relay channel or MARC) and in this context AFR is potentially an attractive option to achieve good performance [10], [33].

Simple wireless repeaters have been used for some time in cellular networks, but relaying techniques are now finding their way into wireless standards. IEEE 802.11 or WiFi systems support simple DF protocols in a peer-to-peer mode, but these cannot be used with wireless access points to provide internet connectivity. The IEEE 802.16 or WiMax standardization group is currently studying relaying technology in working group J. The International Telecommunications Union is currently working on defining new standards activities for future cellular wireless systems under the name IMT Advanced. It seems likely that relays will be one important technology in these systems as cellular operators strive to reduce network infrastructure costs and energy consumption in their networks. The techniques described in Sections III and IV, which introduce interference between sources and relays seem most applicable in scenarios when both have similar transmit power levels. This perhaps makes them more suited for peer-to-peer networks and mobile-to-base station communication links in cellular networks. The MIMO techniques described in Section V could be applied to enhance WiFi, WiMax, and long term evolution (LTE) systems, since all of these standards will soon provide support for MIMO links.

VII. CONCLUSION

In this paper, we have studied the performance benefits of relaying protocols. The conventional relaying techniques can improve link reliability and extend coverage in wireless networks. However, they tend only to be useful in low SNR conditions because of the half duplex constraint. Our review began with NAF and BFNAF protocols which can easily be applied in single relay networks. Concurrent DF protocols use two relays to receive/transmit in turn, thus overcoming the half-duplex limitation. We have also surveyed AFR, DFR, and FFR protocols which can be usefully employed in networks where all terminals have multiple antennas. This survey shows that the techniques described herein can be used profitably in a variety of different scenarios to enhance the benefits of using relaying in wireless networks.

VIII. ACKNOWLEDGEMENTS

I. Krikidis', C. Wang's, and J. S. Thompson's work reported in this paper has formed part of the Delivery Efficiency Core Research Programme of the Virtual Centre of Excellence in Mobile & Personal Communications, Mobile VCE (www.mobilevce.com).

REFERENCES

[1] J. N. Laneman and G. W. Wornell, "Distributed space-time-coded protocols for exploiting cooperative diversity in wireless networks," *IEEE Trans. Inf. Theory*, vol. 49, pp. 2415–2425, Oct. 2003.

[2] J. N. Laneman and D. N. C. Tse, "Cooperative diversity in wireless networks: Efficient protocols and outage behavior," *IEEE Trans. Inf. Theory*, vol. 50, pp. 3062–3080, Dec. 2004.

[3] R. U. Nabar, H. Bölcskei, and F. W. Kneubühler, "Fading relay channels: Performance limits and space-time signal design," *IEEE J. Sel. Areas Commun.*, vol. 22, pp. 1099–1109, Aug. 2004.

[4] A. Sendonaris, E. Erkip, and B. Aazhang, "User cooperation diversity—part I: System description," *IEEE Trans. Commun.*, vol. 51, pp. 1927–1938, Nov. 2003.

[5] A. Sendonaris, E. Erkip, and B. Aazhang, "User cooperation diversity—part II: Implementation aspects and performance analysis," *IEEE Trans. Commun.*, vol. 51, pp. 1939–1948, Nov. 2003.

[6] K. Azarian, H. E. Gamal, and P. Schniter, "On the achievable diversity-multilexing tradeoff in half-duplex cooperative channels," *IEEE Trans. Inf. Theory*, vol. 51, pp. 4152–4172, Dec. 2005.

[7] B. Rankov and A. Wittneben, "Spectral efficient protocols for half-duplex fading relay channels," *IEEE J. Sel. Areas Commun.*, vol. 25, pp. 379–389, Feb. 2007.

[8] Y. Fan, C. Wang, J. S. Thompson, and H. V. Poor, "Recovering multiplexing loss through successive relaying using simple repetition coding," *IEEE Trans. Wireless. Commun.*, vol. 6, pp. 4484–4493, Dec. 2007.

[9] X. Bao and J. Li, "Decode-amplify-forward (DAF): A new class of forwarding strategy for wireless relay channels," in *Proc. IEEE SPAWC*, New York, USA, June 2005.

[10] M. Yuksel and E. Erkip, "Multi-antenna cooperative wireless systems: A diversity multiplexing tradeoff perspective," *IEEE Trans. Inf. Theory*, vol. 53, pp. 3371–3393, Oct. 2007.

[11] B. Zhao and M. Valenti, "Practical relay networks: A generalization of hybrid ARQ," *IEEE J. Sel. Areas Commun.*, vol. 23, pp. 7–18, Jan. 2005.

[12] M. Newton, J. S. Thompson, and J. M. Naden, "Wireless systems resource allocation in the downlink of cellular multi-hop networks," *Eur. Trans. Telecom.*, vol. 19, pp. 299–314, Apr. 2008.

[13] S. Yang and J. C. Belfiore, "Toward the optimal amplify-and-forward cooperative diversity scheme," *IEEE Trans. Inf. Theory*, vol. 53, pp. 3114–3126, Sept. 2007.

[14] I. Krikidis, J. Thompson, S. McLaughlin, and N. Goertz, "Optimization issues for cooperative amplify-and-forward systems over block-fading channels, " *IEEE Trans. Veh. Technol.*, vol. 57, pp. 2868–2884, Sept. 2008.

[15] M. O. Hasna and M. S. Alouini, "Optimal power allocation for relayed transmissions over Rayleigh-fading channels," *IEEE Trans. Wireless Commun.*, vol. 3, pp. 1999–2004, Nov. 2004.

[16] I. Hammerström and A. Wittneben, "On the optimal power allocation for nonregenerative OFDM relay links," in *Proc. IEEE ICC*, Istanbul, Turkey, June 2006, pp. 4463–4468.

[17] M. Dianati, X. Ling, and K. Naik, "A node-cooperative ARQ scheme for wireless ad hoc networks," *IEEE Trans. Veh. Technol.*, vol. 55, pp. 1032–1044, May 2006.

[18] C. Hucher, G. Rekaya, and J. C. Belfiore, "Adaptive amplify-and-forward cooperation," in *Proc. IEEE ISIT*, Nice, France, July 2007, pp. 2706–2710.

[19] N. Prasad and M. K. Varanasi, "Diversity and multiplexing tradeoff bounds for cooperative diversity protocols," in *Proc. IEEE ISIT*, Chicago, IL, USA, June 2004, p. 268.

[20] A. Bletsas, A. Khisti, D. P. Reed, and A. Lippman, "A simple cooperative diversity method based on network path selection," *IEEE J. Sel. Areas Commun.*, vol. 24, pp. 659–672, Mar. 2006.

[21] E. Beres and R. Adve, "Selection cooperation in multi-source cooperative networks," *IEEE Trans. Wireless Commun.*, vol. 7, pp. 118–127, Jan. 2008.

[22] C. Wang, Y. Fan, and J. S. Thompson, "Recovering multiplexing loss through concurrent decode-and-forward (DF) relaying," *Wirel. Pers. Commun.*, vol. 48, pp. 193–213, Jan. 2009.

[23] L. Dai and K. B. Letaief, "Throughput maximization of ad-hoc wireless networks using adaptive cooperative diversity and truncated ARQ," *IEEE Trans. Commun.*, vol. 56, pp 1907–1918, Nov. 2008.

[24] C. Wang, Y. Fan, J. S. Thompson, and H. V. Poor, "A comprehensive study of repetition-coded protocols in multi-user multi-relay networks," revised paper submitted to *IEEE Trans. Wireless Commun.*, Mar. 2009.

[25] Y. Fan, C. Wang, H. V. Poor, and J. S. Thompson, "Cooperative multiplexing: Toward higher spectral efficiency in multi-antenna relay networks," *IEEE Trans. Inf. Theory*, to appear, 2009.

[26] B. Wang, J. Zhang, and A. Host-Madsen, "On capacity of MIMO relay channel," *IEEE Trans. Inf. Theory*, vol. 51, pp. 29–43, Jan. 2005.

[27] Y. Fan and J. S. Thompson, "MIMO configurations for relay channels: Theory and practice," *IEEE Trans. Wireless Commun.*, vol. 6, pp. 1774–1786, May 2007.

[28] O. Munoz-Medina, J. Vidal, and A. Agustin, "Linear transceiver design in nonregenerative relays with channel state information," *IEEE Trans. Signal Process.*, vol. 55, pp. 2593–2604, June 2007.

[29] H. Zheng, A. Lozano, and M. Haleem, "Multiple ARQ processes for MIMO systems," *EURASIP J. App. Signal Process.*, vol 5, pp. 772–782, 2004.

[30] M. Levorato, S. Tomasin, and M. Zorzi, "Coded cooperation for ad hoc networks with spatial multiplexing," in *Proc. IEEE ICC*, Glasgow, Scotland, June 2007, pp. 4746–4751.

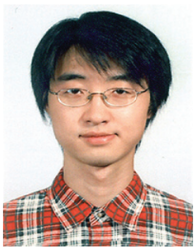
[31] X. Tang and Y. Hua, "Optimal design of non-regenerative MIMO wireless relays," *IEEE Trans. Wireless Commun.*, vol. 6, pp. 1398–1407, Apr. 2007.

[32] Y. Fan, H. V. Poor, and J. S. Thompson, "Cooperative multiplexing in a half duplex relay network: Performance and constraints," in *Proc. ACCCC*, Monticello, IL, USA, Sept. 2007.

[33] D. Chen and J. N. Laneman, "The diversity-multiplexing tradeoff for the multi-access relay channel," in *Proc. CISS*, Princeton, NJ, USA, Mar. 2006, pp. 1324–1328.

Yijia Fan received his B.Eng. degree in electrical engineering from Shanghai Jiao Tong University (SJTU), Shanghai, P. R. China, in July 2003, and Ph.D. degree from the Institute for Digital Communications, University of Edinburgh, March, 2007. His Ph.D. project was fully funded by Engineering and Physical Sciences Research Council (EPSRC), UK. He is currently a Postdoctoral Research Associate in Department of Electrical Engineering, Princeton University. His research interests include signal processing and information theory and their applications in future wireless networks.

Ioannis Krikidis was born in Athens, Greece, in 1977. He received the diploma in Computer Engineering from the Computer Engineering and Informatics Department (CEID) of the University of Patras, Greece, in 2000, and the M.Sc. and Ph.D. degrees from Ecole Nationale Supérieure des Télécommunications (ENST), Paris, France, in 2001 and 2005, respectively, all in electrical engineering. From 2001 to 2002, he served as a Research Associate at the National Capodistrean University of Athens, Athens, Greece and from 2006 to 2007 he worked, as a Post-Doctoral researcher, with ENST, Paris, France. He is currently a Research Fellow in the School of Engineering and Electronics at the University of Edinburgh, Edinburgh, UK. During summer of 2008, he was visiting researcher at the University of Notre Dame, IN, USA. His current research interests include information theory, wireless communications, cognitive radio and secrecy communications. He is a member of the Technical Chamber of Greece.



Chao Wang received the B.E. degree from University of Science and Technology of China (USTC), Hefei, China, in 2003 and the M.Sc. degree (with distinction) from The University of Edinburgh, Edinburgh, UK, in 2005. He is currently a Ph.D. candidate at the University of Edinburgh and participates in the Delivery Efficiency Core Research Programme of the Virtual Centre of Excellence in Mobile and Personal Communications. His current research projects include multiple-input multiple-output (MIMO) wireless systems and cooperative communications.

John S. Thompson received his B.Eng. and Ph.D. degrees from the University of Edinburgh in 1992 and 1996, respectively. From July 1995 to August 1999, he worked as a Postdoctoral Researcher at Edinburgh, funded by the UK Engineering and Physical Sciences Research Council (EPSRC) and Nortel Networks. In September 1999, he was appointed as a Lecturer at the School of Engineering and Electronics at the University of Edinburgh. In October 2005, he was promoted to the position of reader. His research interests currently include signal processing algorithms for wireless systems, antenna array techniques, and multihop wireless communications. He has published approximately 150 papers to date including a number of invited papers, book chapters, and tutorial talks, as well as co-authoring an undergraduate textbook on digital signal processing. He is currently Editor-in-Chief of the IET Signal Processing journal and was a technical programme co-chair for the IEEE International Conference on Communications (ICC) 2007, held in Glasgow in June 2007.

H. Vincent Poor is the Dean of Engineering and Applied Science at Princeton University, where he is also the Michael Henry Strater University Professor of Electrical Engineering. His interests lie in the area of statistical signal processing, with applications in wireless networks and related fields. Among his publications are the recent books MIMO Wireless Communications (Cambridge, 2007) and Quickest Detection (Cambridge, 2009). He is a member of the U.S. National Academy of Engineering, a Fellow of the American Academy of Arts and Sciences, and a former Guggenheim Fellow. He is also a Fellow of the IEEE, the Institute of Mathematical Statistics, and other scientific and technical organizations. In 2005, he received the IEEE Education Medal. Recent recognition of his work includes the 2007 Marconi Prize Paper Award, the 2007 Technical Achievement Award of the IEEE Signal Processing Society, and the 2008 Aaron Wyner Award of the IEEE Information Theory Society.